Information Sciences 323 (2015) 34–47

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Towards non-monotonic sentence alignment

Xiaojun Quan [a,b,1], Chunyu Kit [a,*]

[a] Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China
[b] Institute for Infocomm Research, Agency for Science Technology and Research (A*STAR), Singapore

ARTICLE INFO

Article history:
Received 14 December 2013
Revised 31 May 2015
Accepted 16 June 2015
Available online 25 June 2015

Keywords:
Sentence alignment
Alignment
Non-monotonicity
Bilingual parallel corpora
Bitext
Machine translation

ABSTRACT

All previous works on sentence alignment were founded on the monotonicity assumption that coupled sentences occur in a similar sequential order on the two sides of bilingual parallel corpora (i.e., bitexts), leaving out the non-monotonicity in naturally-occurring bitexts. This paper presents the very first attempt to specifically address this practical issue in sentence alignment, by taking advantage of two observations: (1) an initial (or seed) alignment can be made available using accessible lexical resources and (2) sentences with high affinity in one language tend to have their counterparts with similar affinity in the other. They are incorporated as two constraints into semisupervised learning to formulate a novel and generalized solution for both monotonic and non-monotonic sentence alignment. Our evaluation on real-world data from two remote domains and an end-to-end MT evaluation show that while representative monotonic aligners suffer more severely from a higher degree of non-monotonicity, our approach is able to maintain a stable and competitive performance across the full spectrum of non-monotonicity.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

Automatic alignment of parallel corpora has been a fundamental task in natural language processing (NLP) for two decades. It facilitates many important information retrieval and NLP applications, such as cross-language information retrieval (CLIR) [1,2], statistical machine translation (SMT) [3,4], and bilingual lexicography [5], to name but a few, that demand bilingual knowledge in the form of bitexts (i.e., bilingual parallel corpora/texts as translation of one another in two languages) at various levels of granularity, typically at the sentence and the word level. Sentence alignment to identify correspondences between bilingual sentences plays a pivot role in automatic acquisition of bilingual knowledge, bridging text alignment at the document and the word level. Existing approaches to this issue fall into two main categories, one relying on sentence length and the other resorting to available bilingual lexical resources [6]. The former follows the assumption of a strong correlation between the lengths of coupled sentences, whose validity and effectiveness on aligning monotonic parallel corpora, especially of cognate languages, have been ascertained since the very beginning of text alignment research [7,8]. The latter makes use of word correspondences in a bilingual lexicon (or bilexicon), a kind of more reliable information than sentence length, to infer sentence correspondences. There have also been hybrid methods to combine the strengths and avoid the weaknesses of the two. For example, Moore's aligner [9] takes a multi-pass search procedure to exploit both sentence length and

* Corresponding author. Tel: +852 34429310, fax: +852 34420358.
E-mail addresses: quanx@i2r.a-star.edu.sg (X. Quan), ctckit@cityu.edu.hk (C. Kit).
[1] Performed while a Postdoctoral Fellow at City University of Hong Kong.

http://dx.doi.org/10.1016/j.ins.2015.06.028
0020-0255/© 2015 Elsevier Inc. All rights reserved.

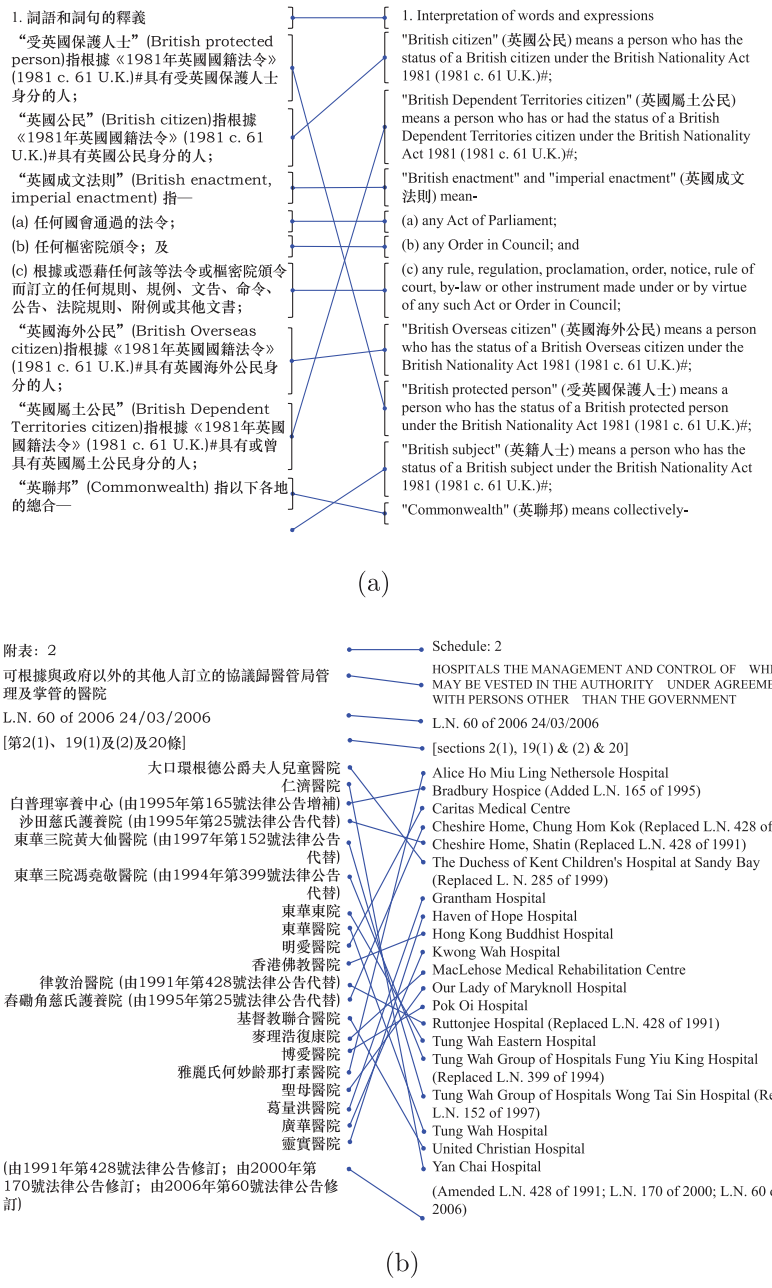

Fig. 1. Examples of non-monotonic alignment excerpted from the BLIS Corpus.

a bilexicon automatically derived from a bitext input. Hunalign [10] is another aligner to utilize these two kinds of information, which backs off, in the case of no lexicon available, to a length-based algorithm to first produce an initial (or seed) alignment for automatic derivation of a bilexicon and then uses it together with sentence length to accomplish the final alignment.

However, all existing approaches depend crucially on the monotonicity assumption, that coupled sentences conform to a similar sequential order in the two languages of a bitext, to such an extent that crossing alignment is in general not entertained [6,11]. As parallel texts available from the web and elsewhere with various characteristics rapidly increase in volume, there has been a necessity to reexamine this assumption, which holds true mostly for strict translation. In fact, the monotonicity prescribed as a prerequisite for sentence alignment cannot always be satisfied, for different languages do not follow exactly the same formula to realize inter-sentential discourse and rhetoric structures. For example, bilingual clauses in legal bitexts are not always organized in the same order. Fig. 1 presents two typical cases of non-monotonic alignment excerpted from the BLIS

Download English Version:

https://daneshyari.com/en/article/391981

Download Persian Version:

https://daneshyari.com/article/391981

Daneshyari.com